Knowledge Distillation : transfert de connaissances entre deux réseaux à l'entraînement

Course organisation

Sessions

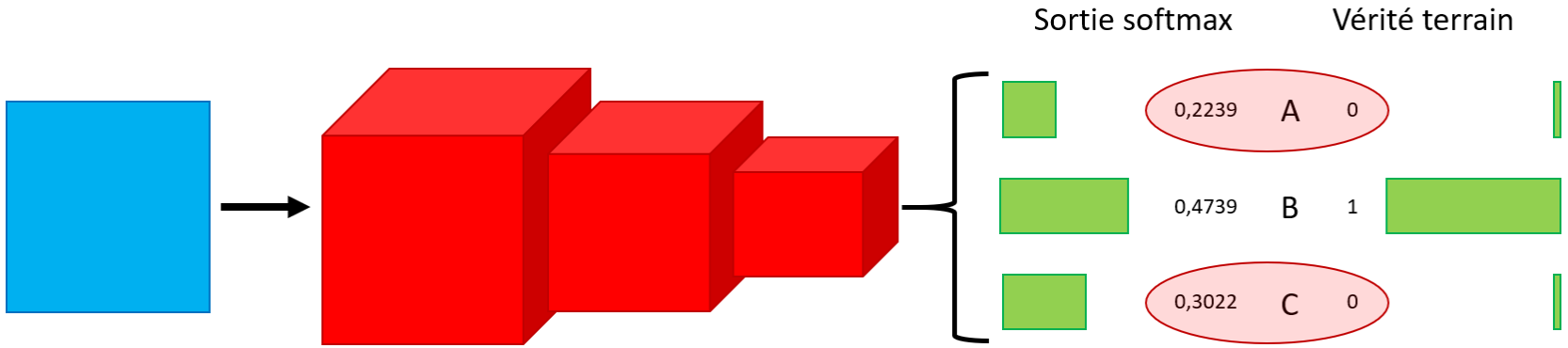
1. Deep Learning and Transfer Learning,
2. Quantization,
3. Pruning,
4. Factorization,
5. Fact. pt.2 : Operators and Architectures,
6. Distillation,
7. Embedded Software and Hardware for DL.
8. Presentations for challenge.

Course organisation

Sessions

1. Deep Learning and Transfer Learning,
2. Quantization,
3. Pruning,
4. Factorization,
5. Fact. pt.2 : Operators and Architectures,
6. Distillation,
7. Embedded Software and Hardware for DL.
8. Presentations for challenge.

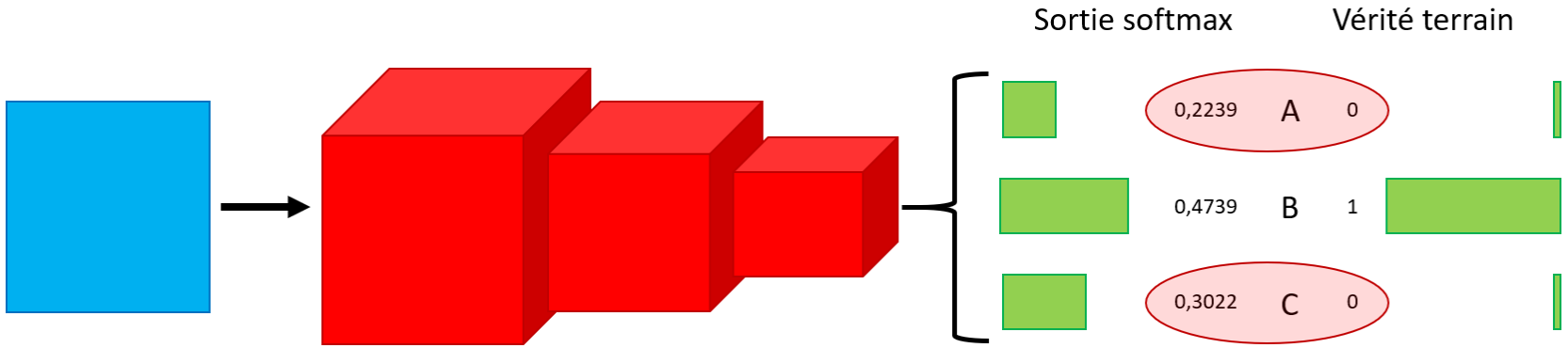
Observation originale



Distilling the Knowledge in a Neural Network, Hinton & al. 2015

- Les valeurs des classes fausses renseignent sur la généralisation du réseau
- Ces labels doux peuvent servir de labels plus pertinents pour entraîner un autre réseau

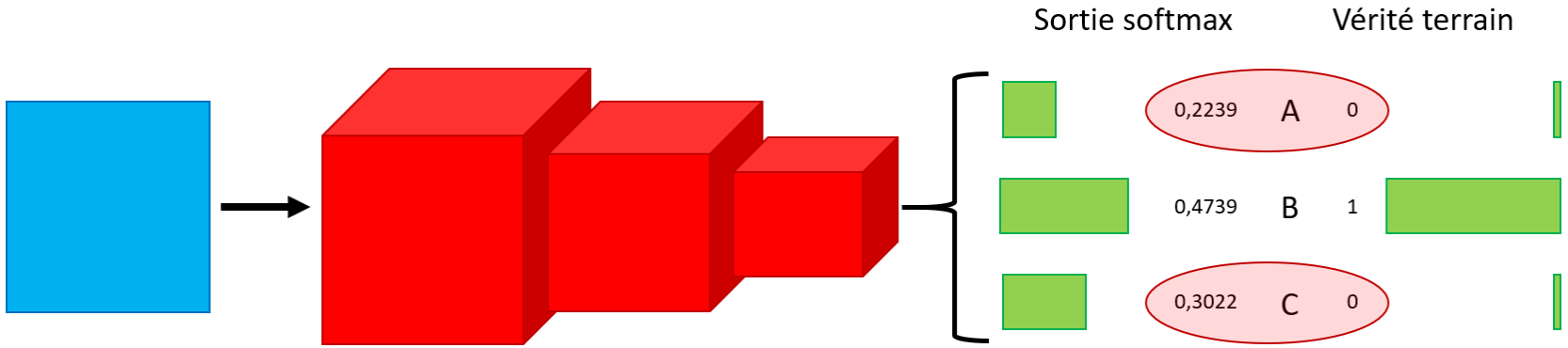
Observation originale



Distilling the Knowledge in a Neural Network, Hinton & al. 2015

- Les valeurs des classes fausses renseignent sur la généralisation du réseau
- Ces labels doux peuvent servir de labels plus pertinents pour entraîner un autre réseau

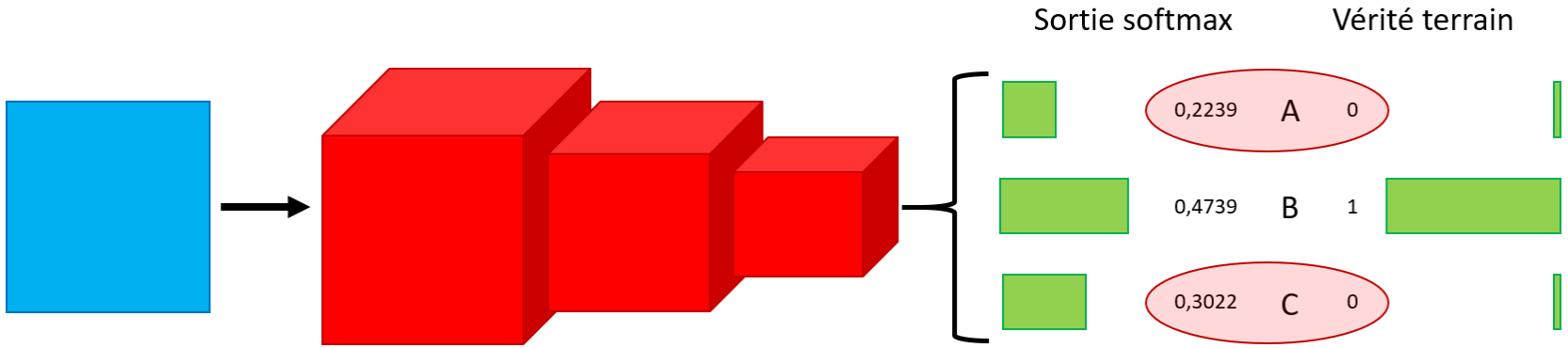
Observation originale



Distilling the Knowledge in a Neural Network, Hinton & al. 2015

- Les valeurs des classes fausses renseignent sur la généralisation du réseau
- Ces labels doux peuvent servir de labels plus pertinents pour entraîner un autre réseau

Observation originale



Distilling the Knowledge in a Neural Network, Hinton & al. 2015

- Les valeurs des classes fausses renseignent sur la généralisation du réseau
- Ces labels doux peuvent servir de labels plus pertinents pour entraîner un autre réseau

Principe de base

Distillation de Hinton

$\mathcal{L}_{KD} = \underbrace{H(y_{true}, P_s)}_{\text{terme supervisé}} + \underbrace{\lambda D_{KL}(P_T, P_S)}_{\text{terme de distillation}}$ avec

$D_{KL}(P_T, P_S) = \sum_i P_T(i) log(\frac{P_T(i)}{P_S(i)})$

et P_S la sortie de l'élève, P_T la sortie du professeur, H l'entropie croisée et D_{KL} la divergence de Kullback-Leibler (ou entropie relative).

Considérons les sorties *softmax* : $q_i = \frac{e^{z_i/T}}{\sum_i e^{z_j/T}}$

- T est choisi égal à 1 pour l'inférence mais supérieur à 1 pour le terme de distillation (les sorties sont alors plus douces)
- Comme les sorties sont alors d'amplitude $1/T^2$, il faut multiplier le résultat par T^2

Principe de base

Distillation de Hinton

$\mathcal{L}_{KD} = \underbrace{H(y_{true}, P_s)}_{\text{terme supervisé}} + \underbrace{\lambda D_{KL}(P_T, P_S)}_{\text{terme de distillation}}$ avec

$D_{KL}(P_T, P_S) = \sum_i P_T(i) log(\frac{P_T(i)}{P_S(i)})$

et P_S la sortie de l'élève, P_T la sortie du professeur, H l'entropie croisée et D_{KL} la divergence de Kullback-Leibler (ou entropie relative).

Considérons les sorties *softmax* : $q_i = \frac{e^{z_i/T}}{\sum_i e^{z_j/T}}$

- T est choisi égal à 1 pour l'inférence mais supérieur à 1 pour le terme de distillation (les sorties sont alors plus douces)
- Comme les sorties sont alors d'amplitude $1/T^2$, il faut multiplier le résultat par T^2

Principe de base

Distillation de Hinton

$\mathcal{L}_{KD} = \underbrace{H(y_{true}, P_s)}_{\text{terme supervisé}} + \underbrace{\lambda D_{KL}(P_T, P_S)}_{\text{terme de distillation}}$ avec

$D_{KL}(P_T, P_S) = \sum_i P_T(i) log(\frac{P_T(i)}{P_S(i)})$

et P_S la sortie de l'élève, P_T la sortie du professeur, H l'entropie croisée et D_{KL} la divergence de Kullback-Leibler (ou entropie relative).

Considérons les sorties *softmax* : $q_i = \frac{e^{z_i/T}}{\sum_i e^{z_j/T}}$

- T est choisi égal à 1 pour l'inférence mais supérieur à 1 pour le terme de distillation (les sorties sont alors plus douces)
- Comme les sorties sont alors d'amplitude $1/T^2$, il faut multiplier le résultat par T^2

Principe de base

Distillation de Hinton

$\mathcal{L}_{KD} = \underbrace{H(y_{true}, P_s)}_{\text{terme supervisé}} + \underbrace{\lambda D_{KL}(P_T, P_S)}_{\text{terme de distillation}}$ avec

$D_{KL}(P_T, P_S) = \sum_i P_T(i) log(\frac{P_T(i)}{P_S(i)})$

et P_S la sortie de l'élève, P_T la sortie du professeur, H l'entropie croisée et D_{KL} la divergence de Kullback-Leibler (ou entropie relative).

Considérons les sorties *softmax* : $q_i = \frac{e^{z_i/T}}{\sum_i e^{z_j/T}}$

- T est choisi égal à 1 pour l'inférence mais supérieur à 1 pour le terme de distillation (les sorties sont alors plus douces)
- Comme les sorties sont alors d'amplitude $1/T^2$, il faut multiplier le résultat par T^2

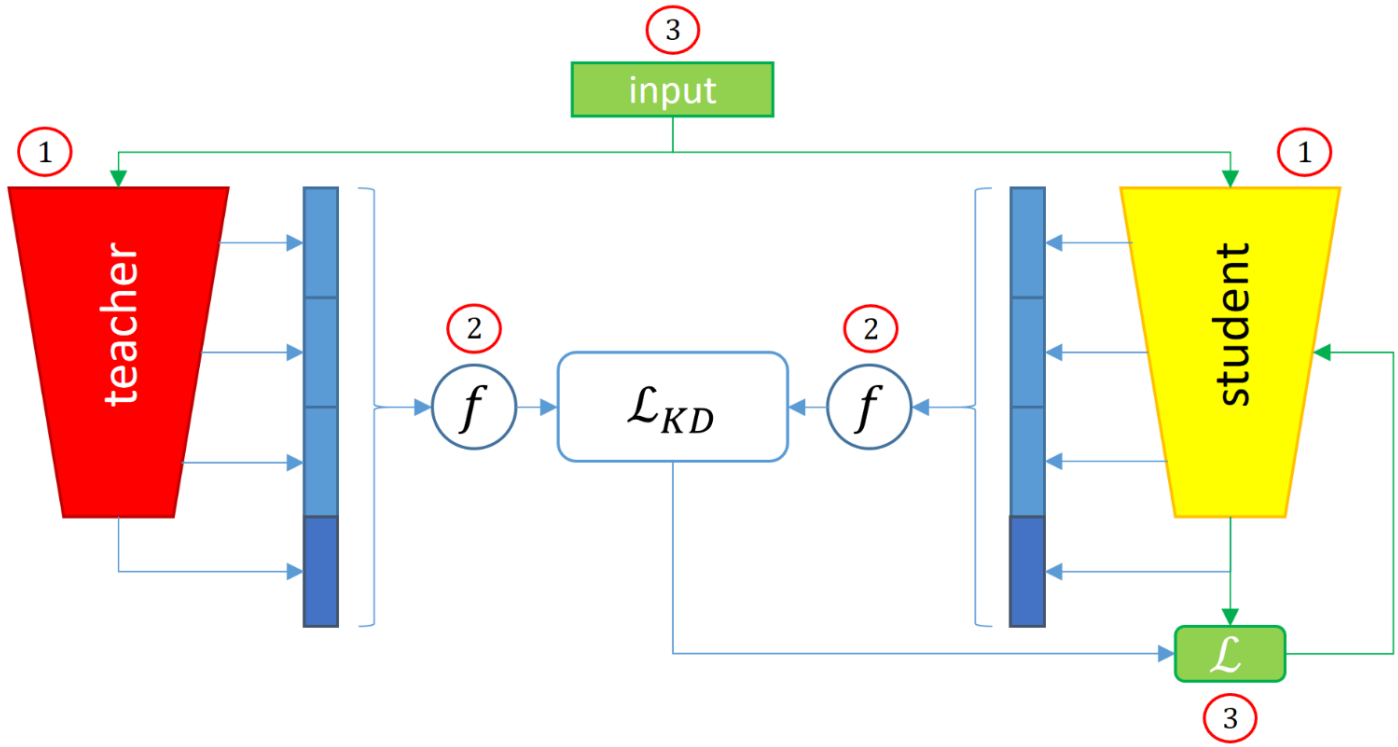
De nombreuses techniques...

Table 5 Performance comparison of different knowledge distillation methods on CIFAR10. Note that ↑ indicates the performance improvement of the student network learned by each method comparing with the corresponding baseline model.

Offline Distillation				
Methods	Knowledge	Teacher (baseline)	Student (baseline)	Accuracies
FSP (Yim et al., 2017)	RelK	ResNet26 (91.91)	ResNet8 (87.91)	88.70 (0.79 ↑)
FT (Kim et al., 2018)	FeaK	ResNet56 (93.61)	ResNet20 (92.22)	93.15 (0.93 ↑)
IRG (Liu et al., 2019g)	RelK	ResNet20 (91.45)	ResNet20-x0.5 (88.36)	90.69 (2.33 ↑)
SP (Tung and Mori, 2019)	RelK	WRN-40-1 (93.49)	WRN-16-1 (91.26)	91.87 (0.61 ↑)
SP (Tung and Mori, 2019)	RelK	WRN-40-2 (95.76)	WRN-16-8 (94.82)	95.45 (0.63 ↑)
FN (Xu et al., 2020b)	FeaK	ResNet110 (94.29)	ResNet56 (93.63)	94.14 (0.51 ↑)
FN (Xu et al., 2020b)	FeaK	ResNet56 (93.63)	ResNet20 (92.11)	92.67 (0.56 ↑)
AdaIN (Yang et al., 2020a)	FeaK	ResNet26 (93.58)	ResNet8 (87.78)	89.02 (1.24 ↑)
AdaIN (Yang et al., 2020a)	FeaK	WRN-40-2 (95.07)	WRN-16-2 (93.98)	94.67 (0.69 ↑)
AE-KD (Du et al., 2020)	FeaK	ResNet56 (—)	MobileNetV2 (75.97)	77.07 (1.10 ↑)
JointRD (Li et al., 2020b)	FeaK	ResNet34 (95.39)	plain-CNN 34 (93.73)	94.78 (1.05 ↑)
TOFD (Zhang et al., 2020a)	FeaK	ResNet152 (—)	ResNeXt50-4 (94.49)	97.09 (2.60 ↑)
TOFD (Zhang et al., 2020a)	FeaK	ResNet152 (—)	MobileNetV2 (90.43)	93.34 (2.91 ↑)
CTKD (Zhao et al., 2020a)	RelK, FeaK	WRN-40-1 (93.43)	WRN-16-1 (91.28)	92.50 (1.22 ↑)
CTKD (Zhao et al., 2020a)	RelK, FeaK	WRN-40-2 (94.70)	WRN-16-2 (93.68)	94.42 (0.74 ↑)

Knowledge Distillation: A Survey, Gou et al. 2020

Évolution de la littérature



1. Quels professeurs et élèves choisir ?
2. Quelles connaissances extraire ?
3. Quel type d'apprentissage ?

Quels professeurs et élèves choisir ?

Professeur

- **Grand réseau**
- Plusieurs réseaux

Élève

- **Réseau plus petit**
- Réseau simplifié
- Réseau quantifié
- Réseau identique

Principalement deux philosophies :

- **Compression** : se servir d'un plus grand réseau pour en améliorer un moins couteux
- **Optimisation** : se servir de la distillation pour améliorer les performances d'un même réseau, ex: *Born-Again Neural Networks*, Furlanello & al., 2018

Quels professeurs et élèves choisir ?

Professeur

- **Grand réseau**
- Plusieurs réseaux

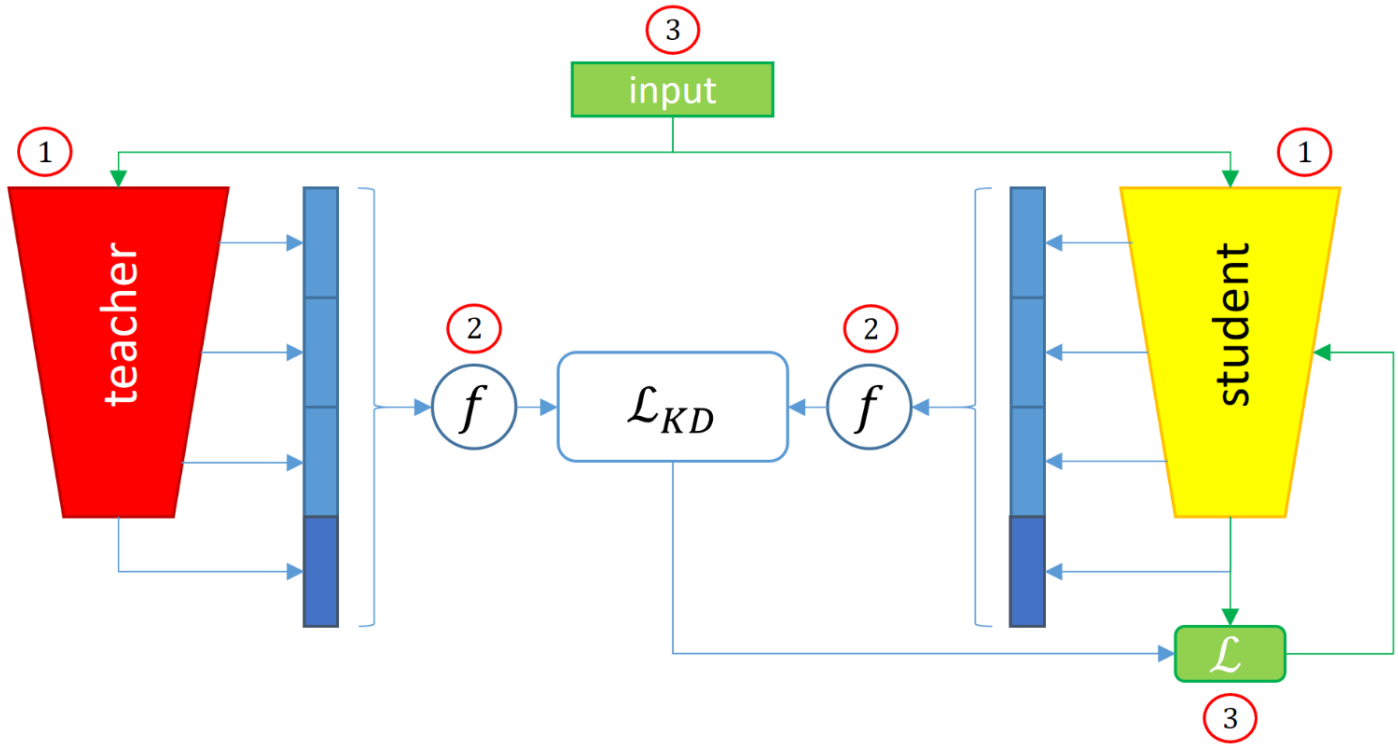
Élève

- **Réseau plus petit**
- Réseau simplifié
- Réseau quantifié
- Réseau identique

Principalement deux philosophies :

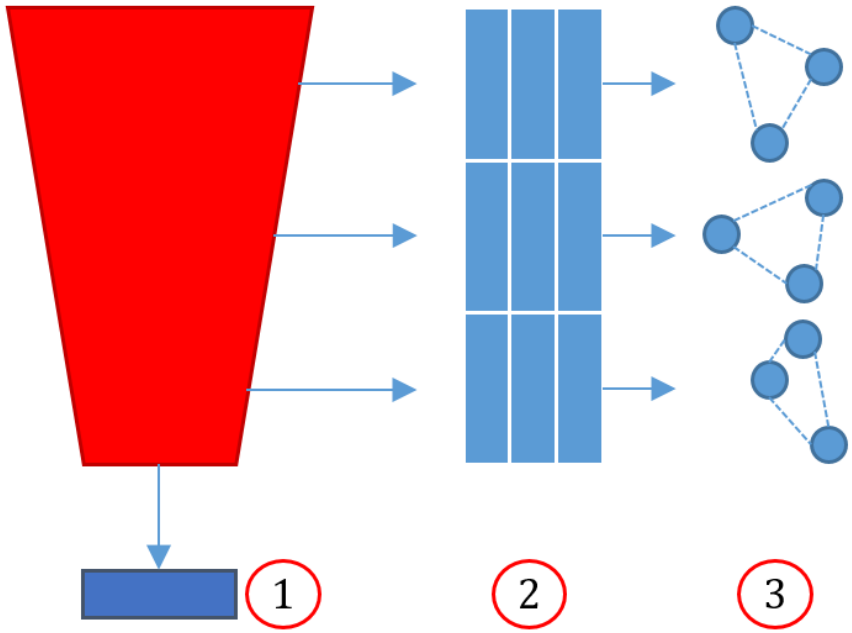
- **Compression** : se servir d'un plus grand réseau pour en améliorer un moins couteux
- **Optimisation** : se servir de la distillation pour améliorer les performances d'un même réseau, ex: *Born-Again Neural Networks*, Furlanello & al., 2018

Évolution de la littérature



1. Quels professeurs et élèves choisir ?
2. Quelles connaissances extraire ?
3. Quel type d'apprentissage ?

Quelles connaissances extraire ? 1/3

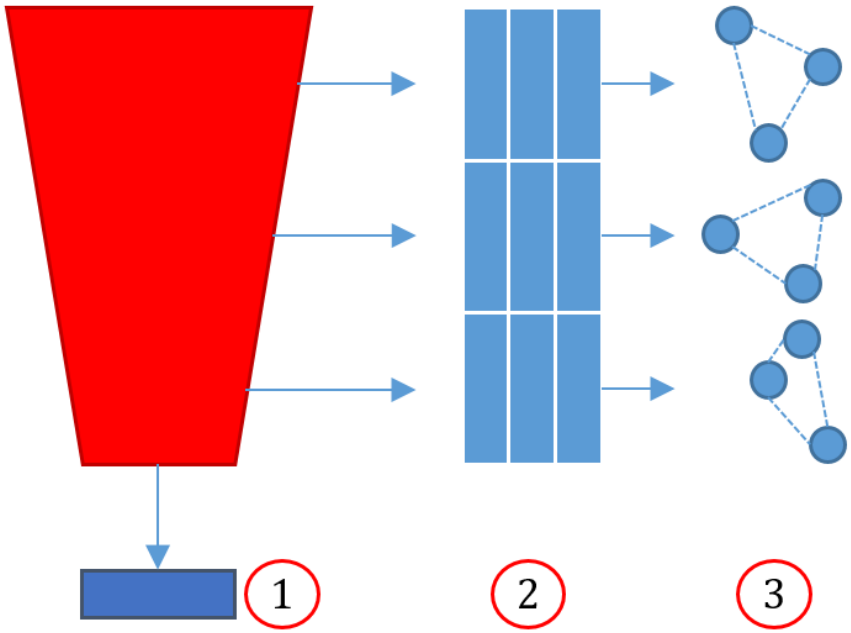


1. Logits de sortie
2. Représentations intermédiaires
3. Relations

Trois articles représentatifs :

1. *Distilling the Knowledge in a Neural Network*, Hinton & al. 2015
2. *FitNets : hints for thin deep nets*, Romero & al., 2014)
3. *Relational Knowledge Distillation*, Park & al., 2019

Quelles connaissances extraire ? 1/3

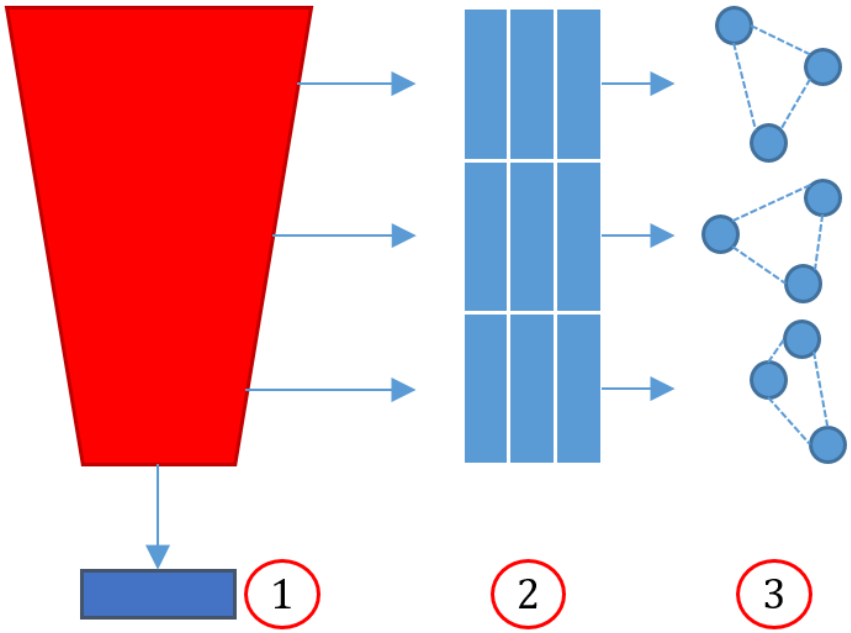


1. Logits de sortie
2. Représentations intermédiaires
3. Relations

Trois articles représentatifs :

1. *Distilling the Knowledge in a Neural Network*, Hinton & al. 2015
2. *FitNets : hints for thin deep nets*, Romero & al., 2014)
3. *Relational Knowledge Distillation*, Park & al., 2019

Quelles connaissances extraire ? 1/3



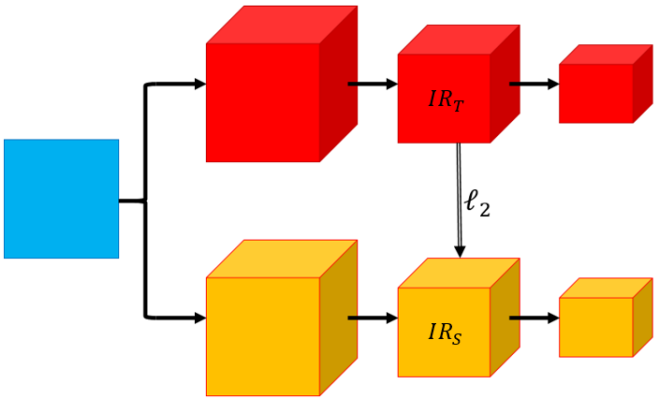
1. Logits de sortie
2. Représentations intermédiaires
3. Relations

Trois articles représentatifs :

1. *Distilling the Knowledge in a Neural Network*, Hinton & al. 2015
2. *FitNets : hints for thin deep nets*, Romero & al., 2014)
3. *Relational Knowledge Distillation*, Park & al., 2019

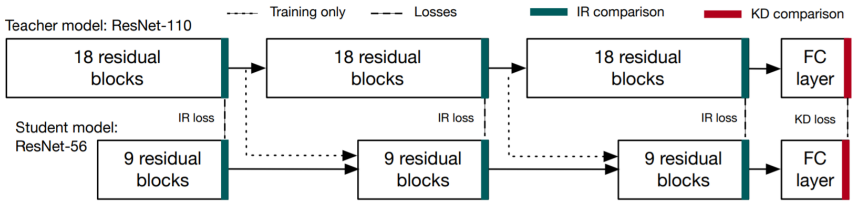
Quelles connaissances extraire ? 2/3

FitNets : hints for thin deep nets, Romero et al., 2014



- Distillation sur les représentations intermédiaires
- $\mathcal{L}_{IR} = \|IR_T - IR_S\|_2$

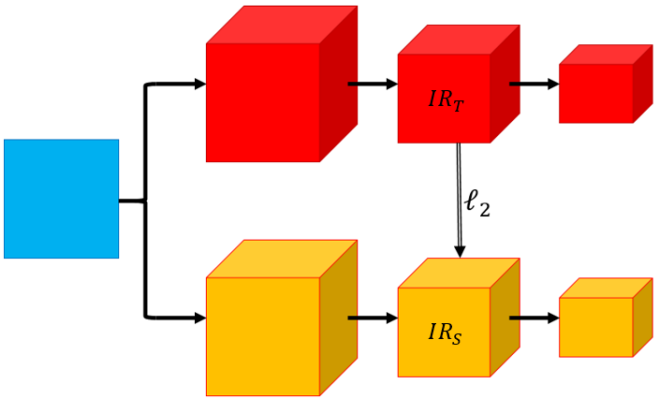
LIT: Block-wise intermediate representation training for model compression, Koratana & al., 2018



- Découpage et isolation de blocs
- Si non correspondance des dimensions : insérer une couche linéaire ou une convolution 1×1

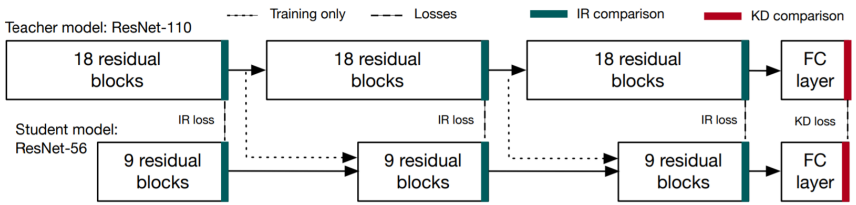
Quelles connaissances extraire ? 2/3

FitNets : hints for thin deep nets, Romero et al., 2014



- Distillation sur les représentations intermédiaires
- $\mathcal{L}_{IR} = \| IR_T - IR_S \|_2$

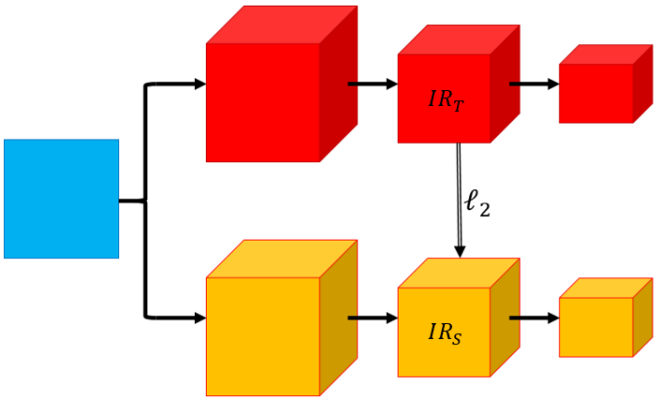
LIT: Block-wise intermediate representation training for model compression, Koratana & al., 2018



- Découpage et isolation de blocs
- Si non correspondance des dimensions : insérer une couche linéaire ou une convolution 1×1

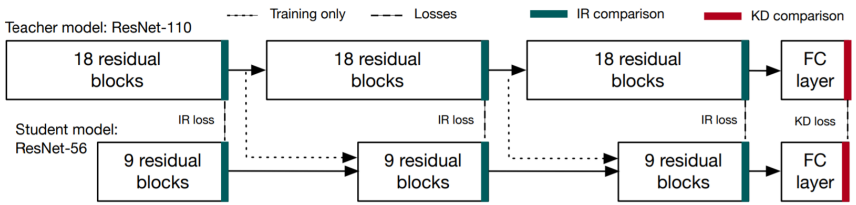
Quelles connaissances extraire ? 2/3

FitNets : hints for thin deep nets, Romero et al., 2014



- Distillation sur les représentations intermédiaires
- $\mathcal{L}_{IR} = \| IR_T - IR_S \|_2$

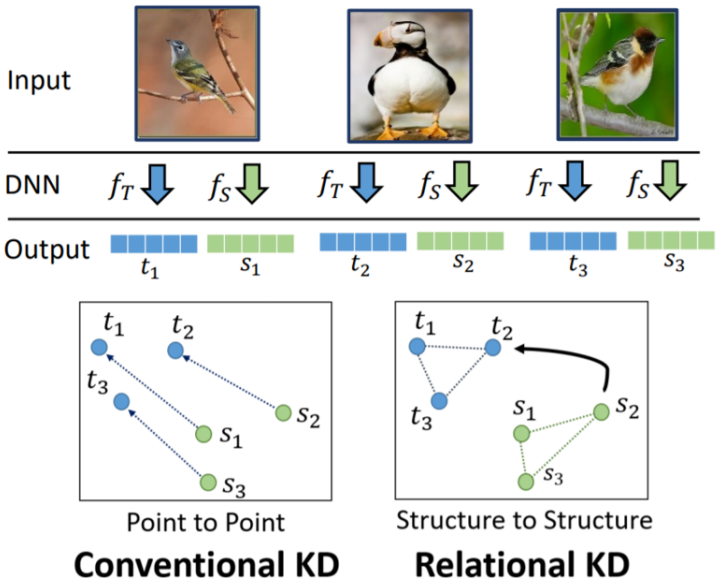
LIT: Block-wise intermediate representation training for model compression, Koratana & al., 2018



- Découpage et isolation de blocs
- Si non correspondance des dimensions : insérer une couche linéaire ou une convolution 1×1

RKD : apprendre comment discriminer les données 2/2

Relational Knowledge Distillation, Park & al., 2019



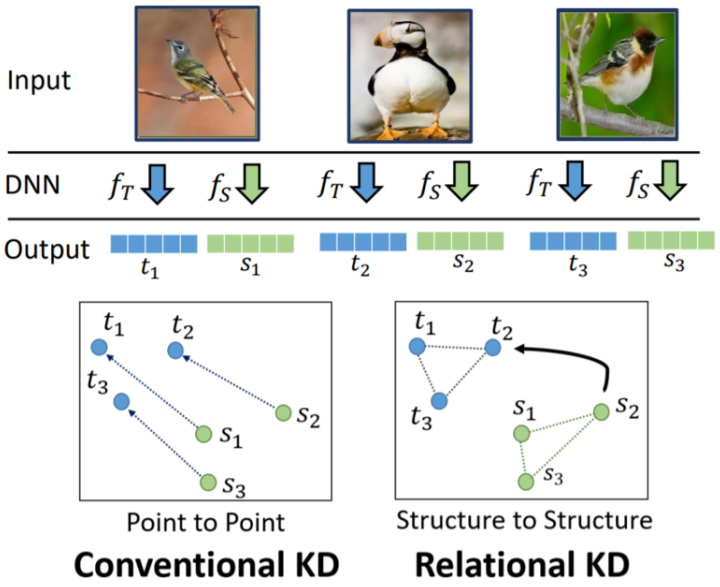
- Abstraction des IR
- Pour chaque batch, norme ℓ_2 entre paires de IR
- On compare ces distances chez l'élève et chez le professeur et on ajoute $\mathcal{L}_{RKD}$ à la loss

Relational Knowledge Distillation

$\mathcal{L}_{RKD} = \sum_{i,j \in \mathcal{X}^N} \ell(\phi(t_i, t_j), \phi(s_i, s_j))$ avec $\phi(t_i, t_j) = \frac{1}{\mu}\|t_i - t_j\|_2$,
ℓ la norme de Huber, alias "norme ℓ_1 douce", μ un terme de normalisation et $\mathcal{X}^N$ le batch d'entraînement

RKD : apprendre comment discriminer les données 2/2

Relational Knowledge Distillation, Park & al., 2019



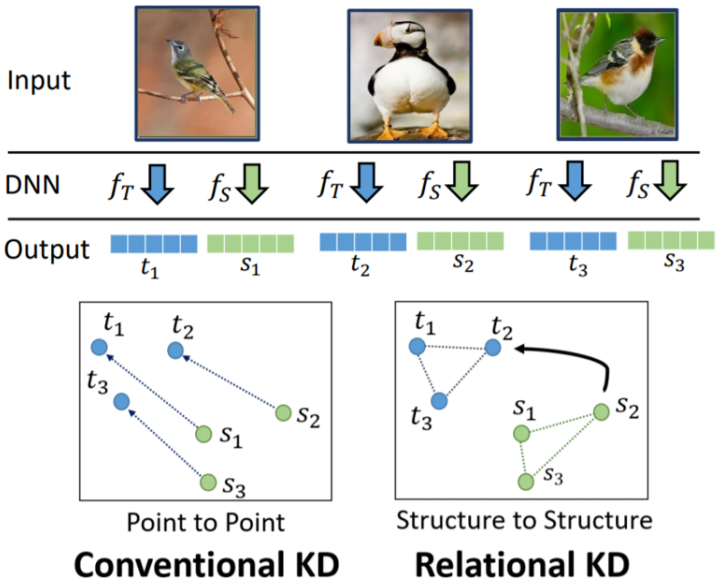
- Abstraction des IR
- Pour chaque batch, norme ℓ_2 entre paires de IR
- On compare ces distances chez l'élève et chez le professeur et on ajoute $\mathcal{L}_{RKD}$ à la loss

Relational Knowledge Distillation

$\mathcal{L}_{RKD} = \sum_{i,j \in \mathcal{X}^N} \ell(\phi(t_i, t_j), \phi(s_i, s_j))$ avec $\phi(t_i, t_j) = \frac{1}{\mu}\|t_i - t_j\|_2$,
ℓ la norme de Huber, alias "norme ℓ_1 douce", μ un terme de normalisation et $\mathcal{X}^N$ le batch d'entraînement

RKD : apprendre comment discriminer les données 2/2

Relational Knowledge Distillation, Park & al., 2019

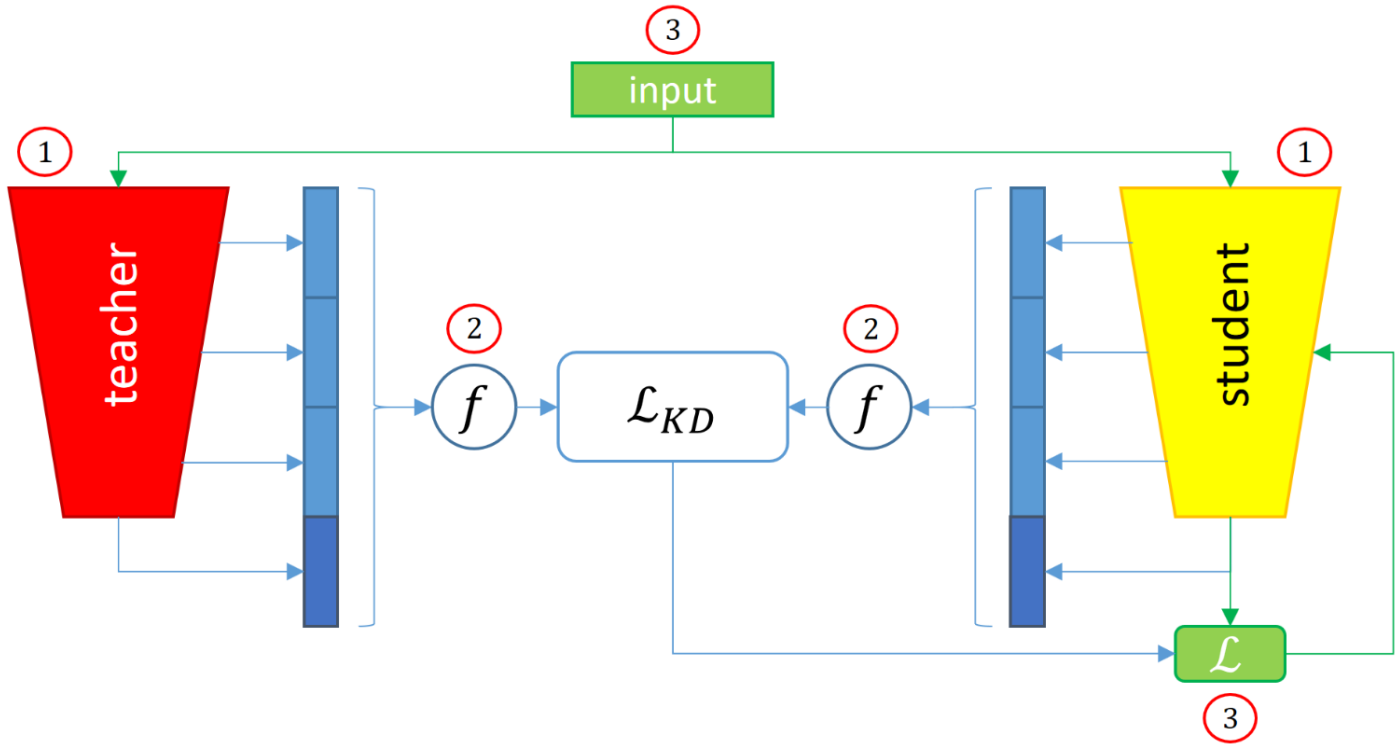


- Abstraction des IR
- Pour chaque batch, norme ℓ_2 entre paires de IR
- On compare ces distances chez l'élève et chez le professeur et on ajoute $\mathcal{L}_{RKD}$ à la loss

Relational Knowledge Distillation

$\mathcal{L}_{RKD} = \sum_{i,j \in \mathcal{X}^N} \ell(\phi(t_i, t_j), \phi(s_i, s_j))$ avec $\phi(t_i, t_j) = \frac{1}{\mu} \| t_i - t_j \|_2$,
ℓ la norme de Huber, alias "norme ℓ_1 douce", μ un terme de normalisation et $\mathcal{X}^N$ le batch d'entraînement

Évolution de la littérature



1. Quels professeurs et élèves choisir ?
2. Quelles connaissances extraire ?
3. Quel type d'apprentissage ?

Quel type d'apprentissage ?

Un domaine très riche...

- Professeur...
  - ... **pré-entraîné** (offline)
  - ... entraîné en même temps (online)
  - ... qui est aussi l'étudiant (self-distillation)
- Données d'entrées...
  - ... **identiques pour le professeur et l'élève**
  - ... différentes (cross-modal distillation, ex: *SoundNet: Learning Sound Representations from Unlabeled Video*, Aytar et al. 2016)
  - ... de synthèse (data-free distillation ou adversarial distillation)
- Entraînement...
  - ... **unique**
  - ... itératif

Quel type d'apprentissage ?

Un domaine très riche...

- Professeur...
  - ... **pré-entraîné** (offline)
  - ... entraîné en même temps (online)
  - ... qui est aussi l'étudiant (self-distillation)
- Données d'entrées...
  - ... **identiques pour le professeur et l'élève**
  - ... différentes (cross-modal distillation, ex: *SoundNet: Learning Sound Representations from Unlabeled Video*, Aytar et al. 2016)
  - ... de synthèse (data-free distillation ou adversarial distillation)
- Entraînement...
  - ... **unique**
  - ... itératif

Quel type d'apprentissage ?

Un domaine très riche...

- Professeur...
  - ... **pré-entraîné** (offline)
  - ... entraîné en même temps (online)
  - ... qui est aussi l'étudiant (self-distillation)
- Données d'entrées...
  - ... **identiques pour le professeur et l'élève**
  - ... différentes (cross-modal distillation, ex: *SoundNet: Learning Sound Representations from Unlabeled Video*, Aytar et al. 2016)
  - ... de synthèse (data-free distillation ou adversarial distillation)
- Entraînement...
  - ... **unique**
  - ... itératif

Quel type d'apprentissage ?

Un domaine très riche...

- Professeur...
  - ... **pré-entraîné** (offline)
  - ... entraîné en même temps (online)
  - ... qui est aussi l'étudiant (self-distillation)
- Données d'entrées...
  - ... **identiques pour le professeur et l'élève**
  - ... différentes (cross-modal distillation, ex: *SoundNet: Learning Sound Representations from Unlabeled Video*, Aytar et al. 2016)
  - ... de synthèse (data-free distillation ou adversarial distillation)
- Entraînement...
  - ... **unique**
  - ... itératif